CREATING HUMAN-LIKE SUBOPTIMAL PLAYERS FOR SCRABBLE®

Mark Richards Andrew Hsi Mykolas Dapkus Daniel Sebastian Pan Thomas Rudwick
Department of Computer Science
University of Illinois at Urbana-Champaign
{mdrichar,hsi1,mdapkus2,dspan2,rudwick2}@illinois.edu

KEYWORDS
Scrabble, Believable Play, Player Modeling

ABSTRACT

Computers have surpassed the level of human play in competitive Scrabble®. However, players who wish to improve their play by practicing against computer programs are often frustrated by the way in which the computer plays. Humans want the option to play against computer players of varying strengths where the differences between levels are natural and human-like. In this work, we show how key components of human Scrabble play can be quantified and evaluated. Based on this data, we create a suboptimal player with features that more closely approximate human play.

INTRODUCTION

Scrabble® is a popular board game inspired by crossword puzzles, where players score points by forming words on a gird. Over 150 million Scrabble sets in 29 different languages have been sold worldwide. Hasbro, which currently owns the rights to the game, estimates that 30,000 Scrabble games are started across the world every hour (Burkeman 2008).

The top Scrabble-playing computer programs are generally superior to human players. Because of the element of chance inherent in drawing the letter tiles, it can take many hundreds of games to accurately determine which of any two players is better. We are not aware of rigorous human-computer comparisons under tournament conditions, but anecdotal evidence suggests that while the very top echelon of human players may be competitive with the best computer players, the AI agents are good enough to beat most humans most of the time.1[1]

Human players often seek to improve their own play by practicing against computer programs. However, people are often frustrated by the mechanical way in which the computer plays. It is easy, for example, to store the full dictionary of legal words in memory and to generate all possible plays. A naïve way to reduce the computer's level of play would be to randomly remove words from its vocabulary or to randomly play only the *n*th best move on a given turn. But this results in uneven and unrealistic play, where brilliant moves are interleaved with sophomoric blunders.

In this work, we seek to quantify and evaluate the effects of three key components of human-level play: vocabulary, play vision, and rack balance. We then use this data to create a suboptimal player whose play more closely approximates the behavior of non-elite human players.

RULES AND BASIC STRATEGY

Scrabble is played on a 15 x 15 grid. Players combine letter tiles drawn from a bag to make plays on the board in a crosswords fashion. The bag initially contains 100 letter tiles. Two tiles are blank and may be used as wildcards in place of any letter. The distribution of the remaining 98 tiles was chosen by the game's inventor, Alfred Butts, based on a frequency analysis of letters from common texts (Fatsis 2001). For example, there are 12 E tiles but only one Q. Each letter tile is assigned a point value; vowels are worth one point each, while the Q and Z are each worth 10 points. Premium spaces on the board can double or triple the value of the letter or word played through them. On each turn, players form one or more words by placing tiles from their rack contiguously in one row or column of the board. At least one tile from the new play must be adjacent to an existing word, except on the first turn when the initial word must be played through the center of the board.

Points are scored for each new word formed on the current turn. All words formed must be found in the Official Scrabble Dictionary, which contains some 178,000 entries. Players start with seven tiles on their racks and replenish their racks to seven tiles after each play, until the bag is empty. When a player uses all seven tiles in one turn, the play is called a bingo and scores a 50-point bonus. Instead of making a play on the board, a player may also choose to exchange one or more tiles from her rack with tiles from the bag. Such an exchange scores zero points.

Basic Strategy

A player's performance in Scrabble depends largely on three factors: vocabulary, play vision, and rack balance. For the purposes of Scrabble, a player's vocabulary consists of all words which he recognizes as legal plays according to the official dictionary. A player need not know the definition of a word in order to use it. In fact, elite players often memorize lists of obscure words which have high Scrabble utility, without bothering to learn definitions.

[1] In a human vs. computer challenge held in Toronto in 2006, the computer program Quackle went 30-4 against top human competition, winning by an average of 125 points per game. In the best-of-five competition between Quackle and the winner of the human tournament, the computer won 3-2, but the results were not conclusive from a statistical standpoint.

But having a large vocabulary is not sufficient. Players must also develop the skill to *see* the opportunity to make high-scoring plays. It takes practice, for example, to see that the rack <AIIMNNV> can be re-arranged to form MINIVAN. We refer to the ability to identify both the right sequence of letters and the appropriate place on the board to make a particular move as *play vision.*

Merely identifying the highest-scoring available word is insufficient for top-level play. Greedily playing the top-scoring play on each turn can lead to a difficult rack down the road (e.g., <IUUVVWY>). Since even the best play for such a rack can be worth only a few points, it might have been better to sacrifice a few points on previous turns in order to get rid of the more difficult tiles. On the other hand, it might be worth sacrificing a few points on the present turn in order to retain high-value tiles such as S or the blank, or to retain combinations of tiles that are commonly found in bingo plays (e.g., AENT). The ability to manage the trade-offs between present and future scoring opportunities is called *rack balance* (Edley and Williams 2001).

Defensive considerations are also important. For example, a skilled player will carefully weigh benefits and risks of a high-scoring play that would open up a triple word score for the opponent.

Computer vs. Human Play

Two top Scrabble-playing programs are MAVEN (Sheppard 2002) and QUACKLE. While the two programs have many similarities, we focus on the latter because it is freely available as an open-source package and was used in our experiments.

With respect to vocabulary, computers have a significant advantage over humans, because it is easy for a program to store the full dictionary in memory. Computers can also easily be programmed to have perfect play vision. On modern hardware, it is possible to generate all possible legal moves (of which there can often be several hundred) in a fraction of a second.

QUACKLE achieves rack balance through a combination of look-up tables for leaves and extensive game tree search. A leave is the set of letters that remain on a player's rack after she makes a play. The quality of a player's leave significantly impacts scoring chances on the next turn. Through analysis of billions of games positions, QUACKLE stores an estimate of the utility of each of the 914,624 possible leaves as a floating-point value. Under QUACKLE's default settings, the legal plays are sorted according equity: the sum of the points scored on the current turn and the estimated value of the leave. Under the default settings, QUACKLE will make the play with the highest estimated equity. This level of play is superior to most human players. Under advanced computer play, the top-scoring plays from the equity analysis will receive further evaluation through extensive game tree search (simulation and evaluation of future possibilities). This analysis will implicitly take into consideration the board configuration, defensive tactics, and the expected value of the remaining tiles in the bag. Human players are not able to store such large lookup tables in their brains or perform the same kinds of brute-force computations but, they can reason at a coarse level about those features explicitly

VOCABULARY

As a first attempt to model the vocabulary of human Scrabble players, we consider a corpus of text from three years worth of articles from the New York Times. After removing words that do not appear in the Scrabble dictionary (the 2006 Tournament Word List), we sorted the words by frequency count. Our belief is that a word's frequency count over a long period of time in a mainstream newspaper would correlate well with the likelihood that an average person would know that word. Table 1 shows a sample of this data. The most frequently occurring words are, not surprisingly, functional words with which English speakers would be quite familiar. Approximately, two-thirds of the legal Scrabble words do not appear at all in the newspaper corpus. The random sample of words shown in the third column show how obscure such words can be. The middle column shows words that appear 20 times. These are primarily words that, while less common, would be recognized as legitimate words.

Top Words	Occurs 20 Times	No Occurrence
the	pumas	antiacne
to	wonky	yays
of	resistor	doggier
and	choosier	blunging
in	ridgetop	eyefold
for	append	triclads
that	brassard	ovately
is	furriers	cooption
said	hotness	acapnia
on	carped	inrushes
by	tenting	inwall
with	jammers	noncom
he	guzzled	maderize

Table 1: Selection from New York Times word list

While such a corpus might help us develop a rough model of the vocabulary of an average English-speaking person, it does not necessarily correspond to the vocabulary of a serious (though not necessarily elite) Scrabble player. As mentioned earlier, Scrabble players understand that certain obscure words are tremendously useful for improving Scrabble play. A widely circulated "cheat sheet" for Scrabble includes lists of all legal two- and three-letter words, short words that include the letter J, Q, X, and Z, words with a high-percentage of vowels, and a list of the most commonly played bingo words. Table 2 shows a sample of some of these words. The fourth column shows a list of words that include the common letters TISANE, plus one additional letter. As players seek to become more competitive players, they will memorize these and other lists.

We conducted experiments to measure the value of knowing certain words. We simulated a round-robin tournament

Twos	Threes	Short Qs	TISANE+?
aa	amu	qi	acetins
ar	dee	qat	cineast
et	gam	qua	fainest
fa	hao	suq	instead
ka	ich	aqua	nailset
mm	jin	qadi	natives
oe	lav	qaid	saltine
oy	naw	qoph	sextain
ut	qis	quag	tisanes

Table 2: Selection of words advanced players may memorize

between five players where each pair of players played 50,000 games. Results are shown in Table 3. All players used QUACKLE's standard equity computation but no tree search. The players differ only in their vocabularies.

Player A knows all and only those words which appeared in the NYT corpus at least 20 times. (We removed obscure two- and three-letter words from the corpus that appeared to have been included only as acronyms.) Player B knows the words in Player A's vocabulary plus all legal two-letter words (60 additional words). We then added the obscure three-letter words (about 500) to make Player C. Player D knows all of the words that Player A knows and also all of the words on the cheat sheet. The cheat sheet includes about 1680 words, and 900 of these do not appear in the NYT corpus. The final player shown in the table, labeled FULL, knows all of the words in the dictionary.

Player1/Player2	B	C	D	Full
A (NYT)	294.1	291.0	283.5	273.5
	344.6	359.8	382.4	430.8
B (NYT+2s)		327.6	320.1	311.0
		352.0	375.3	423.3
C (NYT+2s,3s)			336.0	325.7
			369.7	418.9
D (NYT+Sheet)				347.7
				408.7

Table 3: Comparison of players with limited vocabulary (50000 games).

These experiments yielded some interesting results and show how much of a difference a few key words can make. The player with the basic NYT vocabulary, when pitted against a player with the full lexicon, loses by an average of 157 points per game. Player B beats Player A by 60 points per game, and the only difference is knowing 60 two-letter words! Player D beats Player A by about 100 points per game, while knowing only 900 additional words. The total size of Player D's vocabulary is about 51,000 words. The difference between FULL and Player D is only 60 points, even though Player D knows over 100,000 fewer words. We expect that starting with a corpus-based list and gradually adding key lists of words in this manner will therefore lead to satisfying opponents with human-like vocabularies.

VISION

We next conducted a series of experiments to determine the impact of player vision and leave evaluation. Five players each played ten games against a computer player, in which the computer player always played the first move. For each of these games, the player was allowed to view the list of possible moves and the player selected the highest-equity move for which they knew all of the words formed. This allowed us to remove the factors of inferior human vision and leave evaluation; thus any points lost must be due to vocabulary. By comparing the move selected against the highest equity move for every turn, we computed the average difference in score between the best play and the actual play,

Player #	Avg Pts per turn	Avg Pts lost/turn	Avg Tiles per game	Avg words per turn
Player 1	29.1	7.9	3.6	1.98
Player 2	28.0	8.0	3.4	1.72
Player 3	29.1	6.8	3.6	1.75
Player 4	28.0	12.1	3.4	2.01
Player 5	28.4	10.7	3.4	1.47
QUACKLE	36.7	0	4.1	2.02

Table 4: Human players with perfect leave evaluation and vision.

as well as the average score per turn, average tiles played per turn, and the average number of words formed per move. The results are seen in Table 4. Analyzing these results, we see that even if a player has perfect vision and perfect leave evalution, they will still "leak" points as a result of a stricter vocabulary. Compared to QUACKLE's static player, these players lost more points per turn, scored fewer points per turn, played fewer tiles, and played fewer words with each move. It is worth noting that all five of the players had similar performances, suggesting that they had similar levels of vocabulary.

The same five players then each played three games against a computer player in which no aid was allowed, once again the computer went first. This allowed us to see the impact of

Player #	Avg Pts per turn	Avg Pts lost/turn	Avg Tiles per game	Avg words per turn
Player 1	22.8	14.9	3.3	1.59
Player 2	12.9	22.3	2.9	1.03
Player 3	19.5	15.7	3.0	1.42
Player 4	13.2	27.8	2.6	1.03
Player 5	18.7	17.6	3.1	1.34
QUACKLE	36.7	0	4.1	2.02

Table 5: Human players without any computer assistance

all three factors combined. Again, we compute both the average difference in score per turn and the average points per turn. The results are seen in Table 5. As expected, additional points are lost when players have sub-optimal vision and leave evaluation functions. Comparing the two tables, we can see that the combination of vision and leave evaluation can have a key impact on a player's score. The five players in this study scored fewer points, played fewer tiles, and played fewer words than those who were able to use perfect vision and leave evaulation. Of particular note is the fact that unlike the

	Quackle
Novice	47
Intermediate	6.2
Advanced	3.4

Table 6: Point differential (per game) between human-like leave evaluation strategies and QUACKLE.

previous experiment, the players experienced a large amount of variance in performance. For example, Player 1 scored almost ten points more per turn than Player 2 in this experiment, while their difference in the previous experiment was approximately one point. It is worth noting that this is a small sample size; however these experiments prove that we can apply these tools to analyze sets of games. If we had a larger pool of data, we could use these same tools to process those games and perform a statistical analysis on them.

RACK BALANCE

Our personal experience suggests that novice players do not pay much attention to rack balance. They just try to find and make the highest-scoring move on each turn (and have limited success). Advanced players learn to make the needed trade-offs in order to unload undesirable letters and to retain blanks, S's, and combinations commonly found in bingo plays when the immediate payoff for expending those tiles is insufficient. Because of the sheer number of possible leaves (almost one million), we suspect that even advanced human players will focus their efforts on learning to distinguish between the values of small leaves (1–3 tiles) and not concern themselves with trying to memorize QUACKLE-like values for the larger leaves.

We conducted some limited experiments to compare players who differ only in their rack balance feature. The results are shown in Table 6. In these experiments, the novice player does no leave evaluation. It simply plays the highest scoring play on each turn. This player loses to the standard QUACKLE player by 47 points (Richards and Amir 2007), underscoring the importance of rack balance. The intermediate player has a very basic leave evaluator that is similar to MAVEN's (Sheppard 2002). It has basic evaluations for each single letter and each pair of tiles. It also tries to maintain a balance between consonants and vowels. This player only loses to QUACKLE by 6 points per game. The advanced player is the same as the intermediate player but uses the value computed by QUACKLE for the 500 most commonly occurring leaves. The idea here was to see if memorizing a small number of values would yield a dramatic improvement in rack balance similar to the performance gains we saw in the vocabulary modeling. Clearly, that is not the case. In fact, it appears that the difference between a very crude leave evaluation procedure and QUACKLE's state-of-the-art evaluator is small compared to the differences seen between players due to differences in vocabularies and play vision.

Lexicon	Vision Level	Avg Pts lost / turn	Avg Pts per turn	Avg Pts per game	Avg Tiles per game
Full	1.0	2.4	27.6	339	44.6
Full	0.5	5.3	25.0	311	43.6
Full	0.1	11.3	19.0	248	41.0
Full	0.05	14.6	16.5	219	39.5
Full	0.01	22.2	9.4	140	32.0
Common	Perf.	8.9	21.8	271	43.7
Common	1.0	10.8	19.4	248	41.7
Common	0.05	22.0	9.8	142	33.7

Table 7: Comparison of Created Players

SUBOPTIMAL PLAYER

Given the substantial effects of vision and vocabulary, we created a player model to mimic human behavior by placing restrictions on move selection. For each player, we define two parameters: a numerical skill value that represents how likely the player is to see a move and a file for a vocabulary list. To decide on a move, we begin by sorting all the possible moves based on their score. For each move in the playability list, we consider several key factors. If any of the words are not in the player's vocabulary, the move is discarded. Assuming all words are known, the words themselves are analyzed to determine the likelihood of "seeing" the move. We model vision based on both the number of words formed by a move and the length of each formed words. For example, a move that forms CAT will be very likely to be seen, while a move that forms SERPENT, RE, and TO, is less likely. For each player, 1000 games were run against the default QUACKLE player. In each of the games, QUACKLE made the first move. Table 7 shows statistics for several players made using this model. The "common" dictionary represents a vocabulary containing the words in the New York Times corpus that appeared twenty or more times.

CONCLUSIONS AND FUTURE WORK

We have quantified the effects of vocabulary, play vision, and rack balance in human Scrabble play. Based on the results of our experiments, we have created a suboptimal human-like Scrabble player.

Future work will involve the modeling of these features using a large database of game data from human players.

REFERENCES

Burkeman, O. 2008. Spell bound. *The Guardian* 1.

Edley, J., and Williams, J. D. 2001. *Everything Scrabble*. Pocket Books.

Fatsis, S. 2001. *Word Freak*. Random House.

Richards, M., and Amir, E. 2007. Opponent modeling in Scrabble. *In Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 1482–1487.

Sheppard, B. 2002. World-championship-caliber scrabble. *Artificial Intelligence* 134:241–245.